EMIS Collections and Conference Proceedings
FIZ Karlsruhe

REVIVING THE FREE PUBLIC SCIENTIFIC LIBRARY IN THE DIGITAL AGE? THE EUDML PROJECT

THIERRY BOUCHE

Abstract. In this report we survey results of our efforts to set up a technical environment, a model for external cooperation and interoperability and an organizational framework for a reliable, truly vivid, ever growing and durable digital archive of mathematical publications. A major step forward has been achieved during the course of the European Digital Mathematics Library (EuDML) project which got initial funding from the European Commission from February 2010 to January 2013.

Introduction

In this report we survey results of our efforts to set up a technical environment, a model for external cooperation and interoperability and an organizational framework for a reliable, truly vivid, ever growing and durable digital archive of mathematical publications. A major step forward has been achieved during the course of the European Digital Mathematics Library (EuDML) project which got initial funding from the European Commission from February 2010 to January 2013. I try to give a rather complete picture ot the results of this project in many areas, not focussing on technical issues only.

The general objective of the DML can be summarised as reinventing the free public library in the digital realm. EuDML succeeded in creating a network of a dozen of institutions acquiring selected mathematical content for preservation and access provision, and in making it one single distributed library. The question whether this model will be extended to a much wider scale, or competing models will emerge is entirely open now. However, things have moved thanks to this project, and we can expect exciting developments in this area after a lot of inertia.

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

The paper is organised as follows.
The first section recalls briefly basics of the recent history of the DML project, through the move from the initial vision to some pragmatic rescaling of the EuDML project.
The second section provides an overview of the EuDML project: its objectives, its partners, and the content brought together.
The third section is the main content of this paper. It provides insights about results achieved within the project's three years duration. We highlight the main features of the Web site that is up and running since January 2013. We describe some of the business rules we based the EuDML network on, and expect to build a long-lasting cooperation in the coming months under the name EuDML initiative. Then, the interoperability devices that are currently available are presented. They are designed to ease considerably the use and visibility of the EuDML content from external systems. We finish this section by giving a hint on the various technologies where new software has been developed, tested, or used in an innovative manner within this project.
The last section just sketches some of the challenges that are still to be faced by those who will continue this effort.

Acknowledgement. The work reported here has been partly funded by the European Union through its Competitiveness and Innovation Programme (ICT PSP, Grant Agreement no. 250503).

1. The (Eu)DML Vision

Mathematicians and librarians came up around year 2000 with a vision for a Digital Mathematics Library (DML) that can be summarised by its initial wording by a NSF planning grant that was not followed by much concrete activity (see [24]):

> In light of mathematicians' reliance on their discipline's rich published heritage and the key role of mathematics in enabling other scientific disciplines, the Digital Mathematics Library strives to make the entirety of past mathematics scholarship available online, at reasonable cost, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions.

This vision was instantiated and endorsed by the International Mathematical Union in 2006 [7]. During the first decade of the 21st century, a number of projects were launched around the world (many of them in Europe), which can

be qualified as *local* DMLs. Although these projects met in various occasions, the above mentioned vision didn't foster any cross-border project. David Ruddy [24] and myself [2] suggested that the main inhibiting factor was the overly centralized conception of the foreseen organization. Part of the recent work in this area has thus been to define pragmatic objectives according to a rather bottom-up approach, which can be summarised as follows (see [2, 4]):

> The Digital Mathematics Library should assemble *as much as possible* of the *digital mathematical corpus* in order to
> - help *preserving* it over the long term,
> - make it *available online*,
> - possibly after some embargo period (*eventual open access*),
> - in the form of an *authoritative* and *enduring* digital collection,
> - *growing* continuously with publisher supplied new content,
> - *augmented* with sophisticated search interfaces and interoperability services,
> - developed and curated by a network of *institutions*.

We must stress here that the definition and the need of the envisioned infrastructure is in principle completely orthogonal to the current debate on open access and journal publishing economic models. In fact it should keep neutral to publishing methods and economics as long as the publishing system produces refereed reference texts in some digital format that can be archived and independently delivered through the network of institutions.

Opposite to the trend fostered by electronic publishing towards outsourcing most of fundamental traditional library services to non-public entities [22] (not-for-profit like JSTOR or Portico, or aggressively for-profit like Springer or Elsevier), the above DML vision tries to design the currently vacant function of a distributed library service acting as a reliable back-end to the publishing system, preserving its output, and guaranteeing its availability over the long term.

However, some of the publishing models that got momentum during the last decade do threaten the realization of the DML vision. For instance, proponents of the so-called big deals, as well as the very similar Gold Open Access model where licences are costless to the reader, get most of their value through licensing a vast amount of literature served from the same platform. Duplicating their content is thus going against their business model, even if it's free to download almost everywhere in the world (either because everyone is a subscriber, or because

everyone is entitled open access), but this leaves what might be an important part of the mathematical corpus without curation by a public institution.

Our vision was turned into a proposal to the European Commission and was eventually awarded a grant in the Competitiveness and Innovation Framework Programme, Information and Communication Technology Policy Support Programme, area CIP-ICT-PSP.2009.2.4: "Digital Libraries : Open access to scientific information", grant # 250503, running from February 1st, 2010 to January 31st, 2013 (see [9] for the Commission's view on the project). The proposal was developed under the auspices of the European Mathematical Society (EMS), and specifically its Electronic Publishing Committee.

Compared to previous attempts, this project had two distinct features:

(1) A special attention was given to the fact that the aggregated content would bear mathematical knowledge, hence some provision for mathematical knowledge management was included in the goals.
(2) It was named EuDML although previous EMS lead attempts were under the flag DML-EU.

These features might look marginal at first sight, but it was a radical departure from previous ambitions.

(1) The DML has initially been considered mostly as a mere (and rather simplistic) digital library whose content happens to be mathematical texts. The will of a specific infrastructure would reflect a social feature of the mathematical community (relying heavily on the availability of long-lasting references, and their open accessibility) rather than specific technical needs. We wanted to address the fact that the mathematical nature of the content could be an asset, enabling for instance cross-linking of items based on the mathematical formulae they contain or their mathematical subject according to the MSC, even when they are written in different languages. Therefore, the development and assessment of mathematically savvy technology for searching or handling mathematical content was put at the heart of this project.
(2) While the name DML-EU suggests the European chapter of a (yet to create) global DML (aka WDML: World DML), EuDML was coined to design a European instance of the DML, that nothing prevents to scale beyond its initial boundaries. EuDML being the first such project breaking national borders and bridging distinct local DMLs, the underlying phonetic pun: *Eu* is the half of W ($U = W/2$) was meant to underline

the fact that interconnecting a few projects at the European level would probably amount to half the effort needed to realize this worldwide.

2. Project description

2.1. **Objective.** The EuDML project was explicitly envisioned as a pilot project addressing two challenges that prevented previous attempts towards a global digital mathematics library based on a top-down approach to succeed:

(1) Setting up the technical infrastructure to create a unified access point for the digital mathematical literature hosted by a number of different organizations across various countries.
(2) Defining a cooperation model with a variety of stakeholders that would allow building a reliable global reference library meant to run over the long term, and to be eventually exhaustive.

The two challenges are intimately intertwined as the quality of the technical infrastructure and the array of production and interoperability services provided are the main argument to convince possible partners to join the initiative, which in turn is the best way to enlarge and enrich the content available, thus to reach a critical mass in users.

The technical objective was reached by aggregating a rich metadata repository, and implementing a single access portal for heterogeneous and multilingual collections on top of it. The network of documents has been constructed by merging and augmenting the information available about each document from each collection, and interlinking documents and references across the entire combined library. The most visible outcome targeted by the project was a single access point for the content that was previously dispersed at various places in Europe, with widely varying interfaces and search facilities.

- For users, a website at `eudml.org` with personal work spaces, allowing to search and navigate the collections (see § 3.1).
- For systems, a batch lookup for turning citations into links, as well as a number of interoperability devices to allow automated calls to handled mathematical references (see `project.eudml.org/api`, see § 3.4).

The more political objective has been pursued through various communication channels with a variety of stakeholders, and constitution and consultation of an external Scientific Advisory Board.

The sustainability objective should give rise to the launch of the EuDML initiative by some of the former partners of the project with support from the European Mathematical Society later this year.

2.2. **Consortium.** The project's funded partners are as follows.

- *Overall management & technical coordination:* Instituto Superior Técnico (Lisbon, PT)
- *Scientific coordination:* Université Joseph-Fourier: Cellule Mathdoc (Grenoble, FR)
- Centre national de la recherche scientifique: Cellule Mathdoc (Grenoble, FR)
- University of Birmingham: Computer Science Dpt. (UK)
- Fachinformationszentrum: Zentralblatt (Karlsruhe, DE)
- Masarykova univerzita: Informatique (Brno, CZ)
- University of Warsaw: ICM (PL)
- Édition Diffusion Presse Sciences (Paris, FR)
- Universidade de Santiago de Compostela: Instituto de Matemáticas (ES)
- Institute of Mathematics and Informatics, BAS (Sofia, BG)
- Matematicky Ustav Av Cr V.V.I. (Prague, CZ)
- Ionian University: Informatics Dpt. (Corfu, GR)
- Made Media UK (Birmingham, UK)

From the beginning, the two following institutions were also associated organically to the project.

- European Mathematical Society
- Göttingen university library (DE)

And a Spanish partner had to leave halfway.

- Consejo superior de investigaciones cientificas: IEDCYT (Madrid, ES)

While an Italian partner managed to join.

- Italian mathematical societies (UMI and SIMAI) represented by the Napoli University (IT)

2.3. **Content.** The collections amount to 225,000 unique items (after deduplication), spanning 2,600,000 pages.

Country	Projects	Contributed items
Germany	GDZ Mathematica, ELibM	100,000 items
France	Gallica-Math, NUMDAM, CEDRAM	57,000 items
Czech Rep.	DML-CZ	28,000 items
Russia	RusDML	17,000 items

Country	Projects	Contributed items
Poland	DML-PL	14,000 items
Spain	DML-E	6,400 items
Greece	HDML	3,000 items
Italy	BDIM	2,000 items
Portugal	SPM/BNP	1,300 items
Bulgaria	BulDML	600 items

Out of these, most are retrodigitised (BNP/SPM/IST, BDIM, DML-CZ, DML-E, DML-PL, Gallica, GDZ, HDML, NUMDAM, RusDML) and some are born digital (BulDML, CEDRAM, DML-CZ, DML-E, DML-PL, EDPS, ELibM, NUMDAM).

Selecting content is an important aspect of library collections development. In EuDML we adopted a subsidiarity principle: the project selects partners to be reliable scientific institutions, and relies on them for selecting what out of their holdings will be contributed to EuDML. These decisions are monitored by the Scientific Advisory Board set up by the European Mathematical Society.

EuDML collections are estimated to cover about 6.5% of the whole mathematical reference corpus (estimated to be above 3.5 million items as of 2012). However, the EuDML corpus has some specificities:

- it contains a few books, most of them from the 19th century up to the first half of 20th century;
- it contains a very strong collection of European journals going back to the beginning of 19th century, with many fundamental works;
- the relative coverage of important and long-lasting journal articles is better in the early period when Europe was the centre of the mathematical world, and decreasing with time, as a fast growing number of articles have been published elsewhere.

We have in EuDML 52,156 documents with bibliographic references recorded in the metadata. This resulted in 656,651 individual reference strings. Out of those, 99,282 could be identified as being to EuDML items. Of these, the vast majority (98,000) resolve to articles in journals and proceedings and 1,282 to books.
We infer from these statistics that we succeeded in assembling a non-trivial corpus of reference documents as at least 15% of the citations in EuDML are referring to a EuDML item although EuDML represents only 6.5% of the existing published

documents in mathematics (a given cited item might be counted multiple times here, which is a feature of this analysis: EuDML items are likely to be more cited than the average).

3. Project's outcome

In this section, we first give an overview of the Web interface to the system developed by the project. Then we provide some hints on other results.
As this project intended to create a new infrastructure, with a new way to cooperate for stakeholders involved in the mathematical literature, we first highlight the organization and rules of the EuDML network of partners, then we focus on the results of a more technical flavour.

3.1. **The EuDML Web site.**

The main result of the project, and hopefully the most useful right away is its public Web site `eudml.org` which has been online since more than a year now. However, lots of work has been done in the background so that the current features are worth a visit. With the end of the funded period, features will keep stable for a while (see figure 1).
The EuDML site is intended to be a fully functional digital library with search and browse capabilities. It also allows users to login and enjoy a persistent personalized environment where they can create documents' annotations or personal lists (aka book shelves).
The main page is the entry point for the service. Along with the simple search query interface, it contains basic statistics and some information on EuDML, links to log in or register to the service and links to navigate to other parts of the site.
It is possible to explore the collections using two browsing interfaces: one by subjects (using MSC 2010 classification), and one by journals. Beware however that not all items have been classified according to the MSC, and that there are more item types in EuDML than journal articles (we have books, proceedings published in books and multiple-volume works). At this stage of the system, not all of these can be found readily with the search engine. For instance, you only get to a multiple-volume page from one of its volumes.
The advanced search page allows to perform more sophisticated searches with a boolean combination of positive or negative queries (see figure 2). A unique feature of EuDML is the possibility to search over mathematical formulae written in LaTeX. A formula preview is dynamically generated so that the user can check visually the correctness of the formula. This feature is experimental.

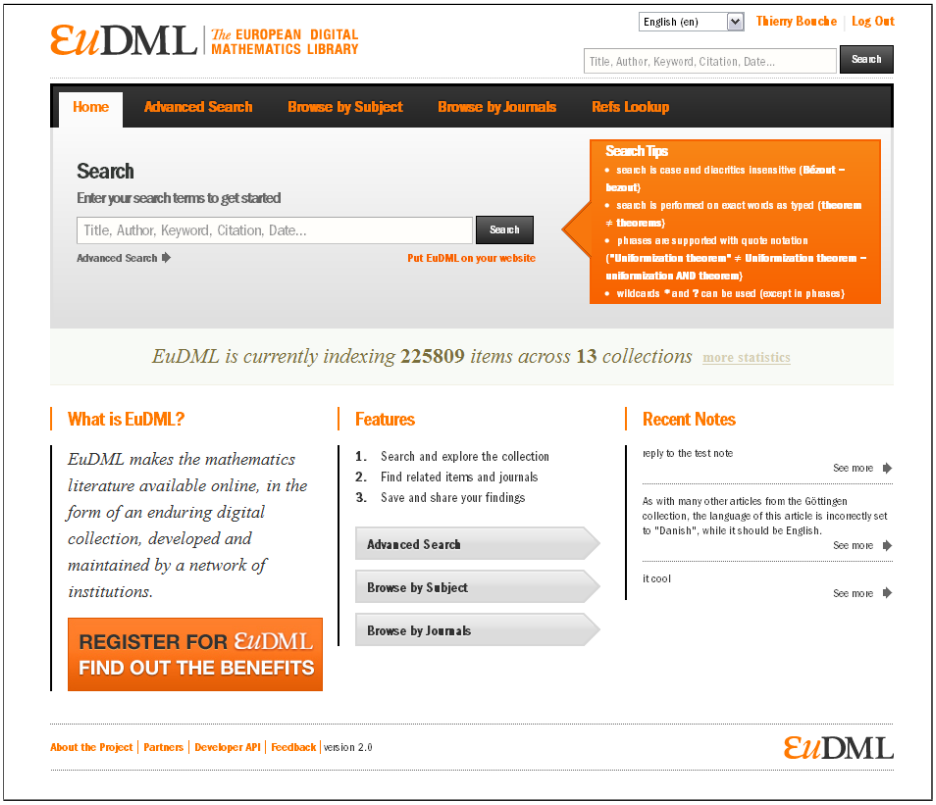

FIGURE 1. The EuDML Web site

When search results are presented it is possible to narrow them using facets. Searched words are highlighted. The math formulae are presented in user-friendly rendered MathML.

The typical landing page for a document contains the most complete display of the metadata known to EuDML for that item, including a full text link at the content provider's site. A number of tools are available to ease further navigation: links from citations, to citing papers, to reviews in zbMath or MathSciNet, looking for similar documents and other relations.

It is also possible to drop a comment, create widgets for embedding the notes on external pages, suggest a correction, add a subject proposition or share the page via social media as well as adding the document to a personal list.

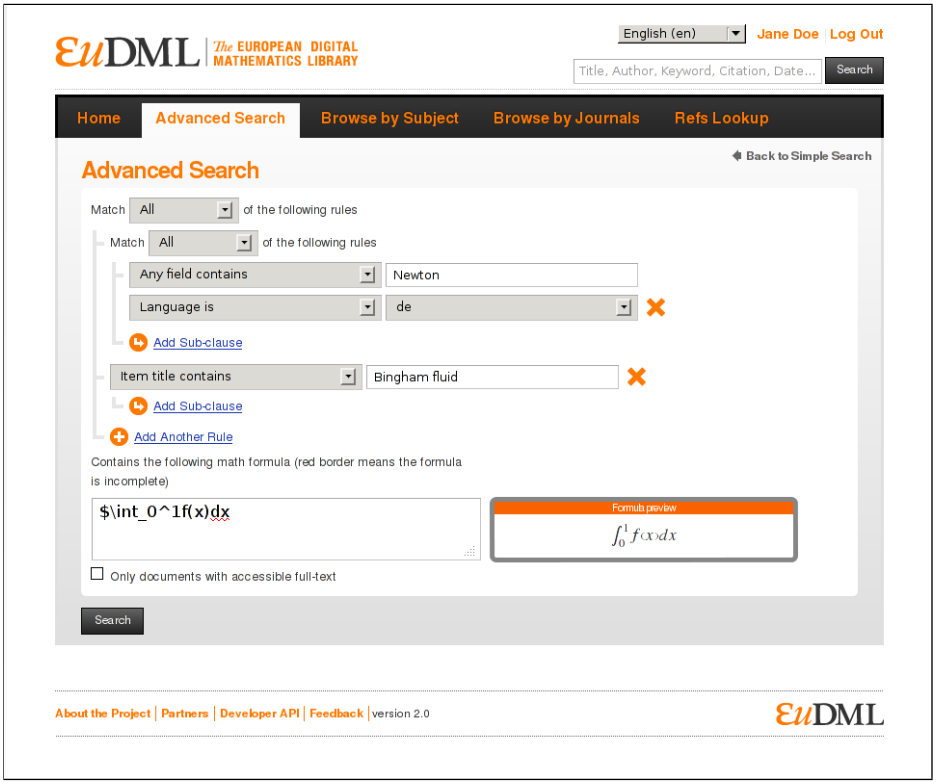

FIGURE 2. Advanced search

One of the main menu item is also unique to EuDML: The Reference Lookup page allows to find a document from a reference citation string, typically copy-pasted from an actual bibliography.

3.2. **Content policy.** EuDML aims to be a long-standing, reliable and open source of trusted mathematical knowledge. This implies to build on firm policy. An outcome of the project is the consensus over the following three conditions for some content to be eligible in EuDML collections.

CP1. *The texts in EuDML must have been scientifically validated and formally published.*

This is needed to ensure that EuDML works as an authoritative library, holding the version of a piece of mathematical knowledge that can be further built upon, and permanently referred to.

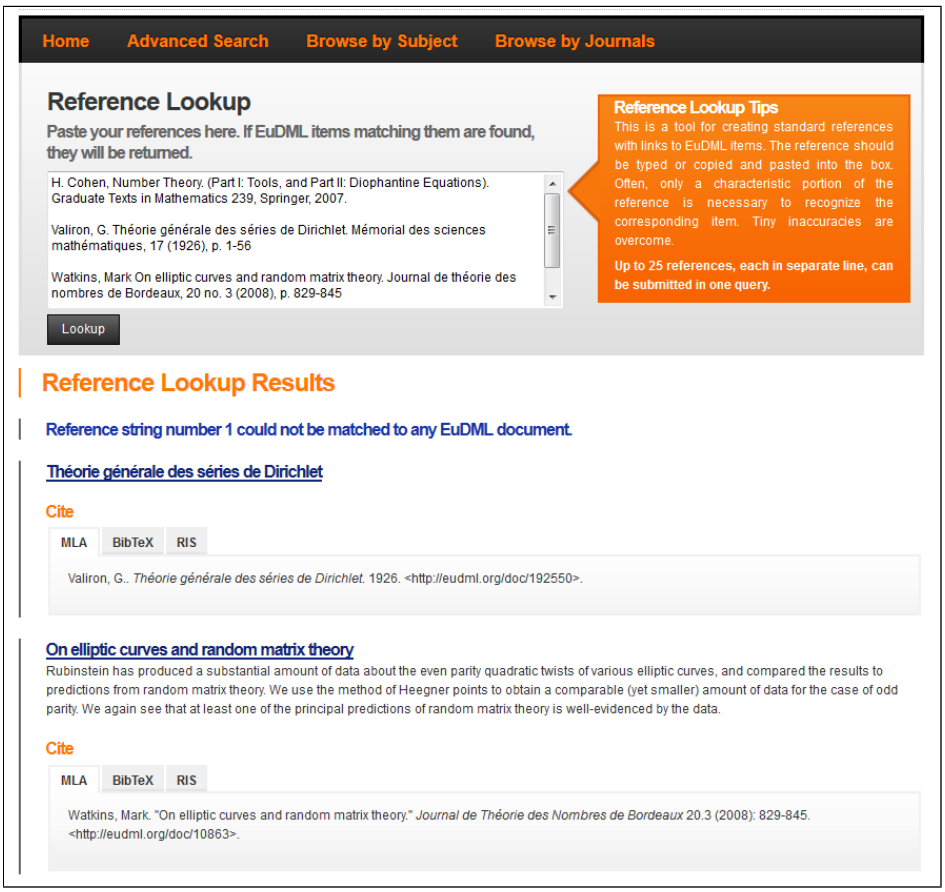

FIGURE 3. The reference lookup.

CP2. *EuDML items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on.*

This is the so-called "moving wall policy" as in general the published items become freed from a pay wall after a certain embargo period (typically ranging from 0, aka open access publishing, to less than 10 years). This eventual open access policy tries to accommodate the fact that not all mathematical publishers can afford to publish everything as open access immediately, but that the value of mathematical knowledge is to

FIGURE 4. A typical item's landing page.

foster new developments in any fields and at any time after publication, so that this should become public knowledge after a not-too-long while (much shorter than current copyright duration, indeed). This policy is strongly supported by the International Mathematical Union as part of CEIC's best practices [6].

CP3. *The digital full-text of each item contributed to EuDML must be archived physically at one of the EuDML member institutions.*

This is for the sake of preserving the mathematical corpus as an enduring collection, which in turn is the only way to secure its online availability over the very long term.

We noticed during this project that these rules are strong and will limit our ability to reach an exhaustive mathematical corpus (the Elsevier archived 'primary mathematical journals' [8] that have been recently released as open access would comply with the first two of them, for instance, while not all project Euclid journals would comply with the second one). However we felt that these rules ensure that the system we built is on a sound base, and that what has been achieved so far cannot be reverted by some external fortune.

Our Scientific Advisory Board commented on these policies at the very end of the project and suggested that we relaxed them somehow, in order to maximize the eligible content. Tweaking these policies so that EuDML is as inclusive as possible but not just a loose index of untrusted mathematical papers on the Web was a challenge of this project. We thought that we should first have a strong perpetual content base before trying to accommodate with looser scenarios.

3.3. **External Cooperation Model.** Based on the above content policies, we drew a model of EuDML operation that will inform the design of the EuDML Initiative. It is based on a consortium of EuDML core members being scientifically and organizationally strong not-for-profit institutions that take care of the system's activity, maintenance, and of the collections both in terms of preservation and eventual open access provision. This gives rise to a network structure relying on a core set of internal partners providing content and technology. A second tier is foreseen to allow participation of associated partners that, for instance, use some services from a sponsoring first tier partner to access the network. A typical scenario for this is that of a content partner willing to contribute collections but without the skills or resources to comply with the interoperability requirements, thus engaging with one of the core partners that would serve as entry point for them. This structure is already active for some of the project's content. A third tier would consist of external content partners (typically publishers) that

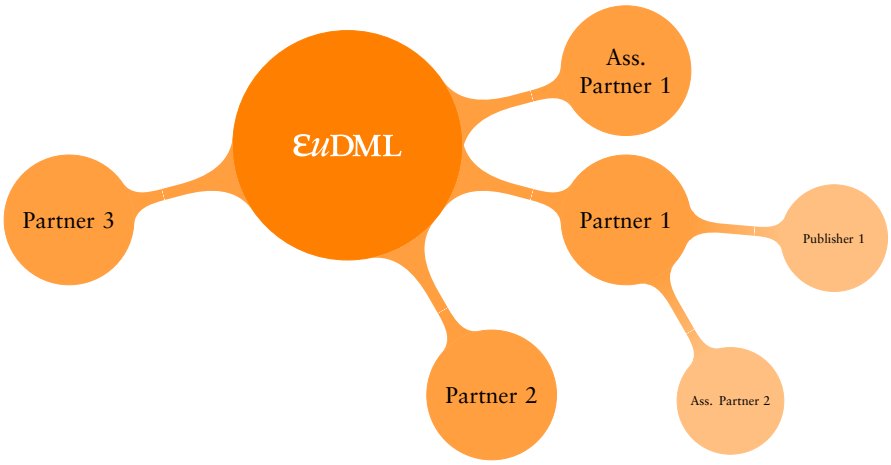

FIGURE 5. The EuDML network

contribute, possibly directly, metadata to the project, but need to transfer their collections to one of the core members in order to comply with our content policy CP3.

The EuDML content members should

- be aligned with the project's goals,
- keep committed over the long term,
- select collections to be contributed to EuDML on sound scientific grounds,
- develop a preservation policy for the full-texts,
- acquire new items in a timely manner (retrodigitisation or direct from publishers),
- sort out rights and licences of contributed collections,
- take care of data and metadata curation,
- manage communication with the central registry.

The EuDML technical members should

- be aligned with the project's goals,
- keep committed over the long term,
- manage communication with the content members,
- run and maintain parts of the system's infrastructure,
- develop new services as the need emerges and to the extent their resources permit.

External partners are expected to contribute to the EuDML Initiative using our interoperability model. We identified the following typical scenarios.
External content partners should

- adhere to the project's goals,
- select one content member (aka local DML center: LDC) as entry point to EuDML,
- set up transfer and update mechanisms for new items,
- determine the moving walls' durations,
- license at least one LDC to store transferred files for ever.

External technical partners should

- adhere to the project's goals,
- sign non-disclosure agreement of data they could get hold of for their technical work,
- develop technology over subsets of the corpus and make it available to the project,
- provide technology to the project preferably under open source licenses.

External linking partners should exploit the linking opportunities delivered by the project to enrich content and user experience while searching, browsing, or accessing the reference mathematical corpus.
The Scientific Advisory Board, in line with its previous comment on policies, advocated for a "second level partnership" with relaxed implications. By publishing these results, we hope to get more feedback from the community on the operation model we invented.

3.4. **Interoperability Model.** In order to enable many interoperability scenarios, a number of tools have been developed and deployed. The goals pursued are, on one side to make it easy to contribute new content to the EuDML system, and on the other side to offer many useful ways of exploiting the EuDML content, or creating specific views for different communities.

Contributing content to EuDML. The preferred mechanism to contribute content to EuDML is to set-up an OAI-PMH server to export XML metadata structured according to the EuDML schema version 2.0, providing the mandatory elements and tagged according to the best practices that are specified on its website [13]. These specifications have been designed so as to impose minimal technological barrier to content providers yet to enable the transfer of highly detailed and accurate metadata. Many publishers already export JATS files to interoperate with services such as Portico, JSTOR, PubMed Central, etc. To help content providers

tweak their EuDML metadata, we provide them with an online validation tool [14], which is also applied in the ingestion workflow. This model is the preferred one as it requires almost no work on EuDML side to ingest or update new content, thus will be available after the project funding expires.

For those content providers who are not able, or not willing, to export metadata prepared according to our recommendations, we developed a number of transformations from various flavors of OAI-DC, which are performed on-the-fly at ingestion time.

For those content providers who cannot set-up an OAI-PMH server delivering the expected metadata (missing mandatory elements, e.g.) but do have it in some supported format, it is in many cases possible to harvest files through FTP, then run on-the-fly transformations, so that the ingestion process looks transparent to the central system.

Finally, we have started to build the second tier of the EuDML network, where a EuDML partner "sponsors" an associated partner by getting hold of its relevant metadata, doing the necessary transformations, and posting them to EuDML from its OAI-PMH server. Mathdoc had this role for the collections from GDZ, DML-E for instance.

After the initial pilot period of EuDML, it is envisioned that publishers should contribute to EuDML in this way: selecting a EuDML member that would host a copy of their content, and make it available to EuDML (in fact, this is the scheme already in use for most of the digitised collections).

Our impression at the end of the project is that this model works pretty well. Some "second tier" partners at the beginning of the project (DML-CZ, e.g.) are now "first tier". It is however an open problem to tell to what extent this model can scale from the current 200+ thousand items to the 1.5 million items in the scope of DML estimated to exist currently in digital form, or even to the 3.5 million mathematical items published worldwide since Euclid. Probably the main barrier here is not technical, it was already addressed in previous sections. But there are lots of small collections out there that would be eligible to EuDML but wouldn't enter into one of the above tiers. Examples of such collections are numerous. Typically this can be a large digital library holding a very tiny portion of mathematical content for which no dedicated work or resources can be allocated, or a very amateur digital library set up by a small group of unskilled people, the extreme version of this being an author's own works digitised or collected on his own web page (the IMU called all mathematicians worldwide to do so). In fact these collections would require some pro-active action from a EuDML partner to be exploitable by EuDML. It is not obvious to tell what

portion of the content that could be available that way would ever reach EuDML through another path, but it is clear that breaking this barrier would enlarge considerably the content. This challenge was not addressed in this project but should be investigated later. As always, the low hanging fruits were caught first, and resources needed grow exponentially with height!

External interoperability devices. While a smooth ingestion procedure is the guarantee for EuDML to register an up-to-date critical mass of quality metadata, fuelling powerful discovery services and a rich user interface, external interoperability is needed to allow third parties to enrich their services thanks to the availability of collections in the EuDML system. This in turn provides more visibility to and more usage of the EuDML collections.
We developed specific tools for targeted scenarios of machine interaction with the EuDML corpus.

(1) Batch download of public elements in descriptive metadata is available through the EuDML public OAI-PMH server [10]. In order to maximize interoperability, three formats are supported: basic OAI-DC, Europeana semantic elements [16], and EuDML schema [13]. Apart from some sensitive data that can have been contributed to EuDML under the condition that it is not re-served (author's email addresses, copyrighted full-texts, e.g.), all information that EuDML harvested or created (EuDML Ids, links to other databases, e.g.) are exported under the last format. It is thus also a way for EuDML content providers to get back the project's added value for their own sake.
(2) Machine query the EuDML database with Opensearch [11] using Contextual Query Language syntax. This would allow a third party to automatically query the EuDML database and present EuDML hits together with other sources, for instance.
(3) Machine calls to some EuDML functions through REST services [12]. These services have been tailored for various needs, and should probably evolve depending on feedback or as new needs emerge.
    (a) The Batch Ref service allows an external party to upload a reference list with citations of mathematical documents, and get back the identifiers of matched EuDML items. This is a critical added-value for a reference library as this allows many stakeholders dealing with mathematical references to enhance their assets by adding links to the full-texts.

(b) The Reverse Ref service makes it possible to find all EuDML items citing a given EuDML item. This service was an explicit request of a putative content provider in order to get an added value from participating in EuDML, as it would generate more valued links to their assets.
(c) The Similar Items service makes it possible to use the EuDML website's "Find similar documents" feature from a distance.
(d) The Batch Ids service allows third parties knowing one Id for a given item to query the EuDML databases for all Ids pertaining to this item known to the database. It turns EuDML into a mathematical hub connecting relevant databases. Together with the All Pointers service, it opens new pathways in the mathematical corpus.
(e) Finally, the Metadata via REST service makes it possible to download an item's internal metadata in two XML formats.

(4) Embed some EuDML data or query form as a widget in a Web page. For instance, users can monitor their EuDML activity or add some dynamic view on EuDML in their Web site.

These tools open a wide range of possible applications, from adding the EuDML corpus to an external search engine to enriching existing content with deep links to EuDML.
Producing Linked Open Data and creating a SPARQL end point was considered during the project, but the technology didn't seem mature enough for a production system, real-world application still lacking to exhibit a clear benefit within the short time frame for development. We also have in principle the possibility to set-up a full-text hub as the central system does store quite a lot of full-texts from EuDML content providers, in quite many formats (original PDF, extracted text with or without math as MathML or LaTeX, accessible formats) and we also have licence declarations from the content providers whether these texts can be used internally only for indexing, or can be re-served openly. However, these services were not developed in this project.

3.5. **Technical results.**

3.5.1. *Metadata.* One of the most basic yet non-trivial challenge in the project was to agree on a common metadata format, as each partner had its own, and stood with quite varying background, technical as well as in terms of the community they belonged to.

After a rather involved discussion, we adopted the NLM Journal Archiving and Interchange Tag Suite as the basis for EuDML metadata storage and exchange, which became an NISO standard during the course of the project [20].
To handle the extra content (monographs, edited books or proceedings and their chapters, multiple-volume works and their volumes), we created a new XML schema that defines a specific superstructure and relies on the standard article elements for all shared concepts.
This metadata format supports all EuDML item types so far, and still leaves room for storing improvements such as structured XML full-text, or multiple versions of the same citation. As we store the best available metadata, it is easy to generate simpler schemas such as OAI-DC or Europeana semantic elements.
The metadata is harvested and mostly transformed on-the-fly to JATS by the REPOX harvest manager developed by our partner in Lisbon, mostly in connection with the digital programme of the Portuguese National Library and Europeana.

As the project was also an occasion to clear licenses and copyright for the contributed content, we can report the following.

- We estimate that 97% of full-texts as PDF files are openly accessible from their providers while only 10% are old enough to be public domain.
- The metadata as available from EuDML OAI-PMH server is entirely freely reusable according to either CC0 (public domain) or CC-BY (attribution) Creative Commons licences.
- For full- texts, the situation is somewhat more complex:
  - 135,000 items have some sort of text-only full-text that is usable for indexing purposes, coming from text OCR or PDF extraction,
  - 170,000 items are available for project internal processing such as re-OCR to get math formulae or as test-bed for whatever enhancement a partner could try (most of them are scanned PDFs, but some are born digital),
  - the PDFs of 105,000 items could be re-served after some processing such as adding text or math layers to an image PDF. However, only 10,000 files have been processed with Maxtract [1] and are currently served in some new format generated for print-disabled users.

3.5.2. *Productivity tools.* A number of productivity tools were produced in the course of the project. They are usually Open Source software or libraries. We provide some live demos on the project's Web site [15].
Here is a list of services running in the background or enabling some of the Web site features.

- Metadata enhancements (automated tagging refinement such as author names or keywords splitting);
- On-the-fly conversion from TEX encoding of formulae to MathML (based on Tralics [19]);
- EuDML reference matching, zbMath matching;
- Item metadata merging: we had some 2,000 items duplicated from different partners: we created a single record for them.
- Public demo website with presentation MathML based display of formulae (using MathJax [5] as a fall-back)
- Experimental formula search (based on Brno's WebMIaS [25])
- Experimental similarity computation (based on Brno's Gensim [23])
- Experimental production of accessible formats of mathematical texts (based on Birmingham's Maxtract [1])
- Web 2.0 features and annotation module
- Service interfaces (Opensearch, OAI-PMH, REST API)
- More mathematical knowledge generated and stored in XML records through
  - MSC and English keywords acquired from zbMath
  - Text+MathML extraction from born digital PDF (using Birmingham's Maxtract [1])
  - Text+MathML extraction from image PDF (using InftyProject's InftyReader [26])

All these bits and pieces were integrated and made to work together by the team at ICM Warsaw where the central system is running.

4. Open questions and future work

4.1. **Content acquisition.** The main point to users is the content: it's nonsense to learn the interface of one more search engine if it covers less than 10% of the whole corpus. On the other hand, there are many mathematical texts that do exist digitally, and are freely accessible on-line, but can't be located easily from mainstream search engines or even dedicated reviewing databases. A large part of the retro-digitised corpus is hidden because it lacks full-text and can only be searched using scarce metadata (or almost non-existent metadata, as for PDFs linked from hand-made HTML pages). The fact that users start using EuDML to locate and refer to papers from DML-PL or GDZ shows that we shifted the state-of-the-art in this respect. However, to succeed, we need to cover much more of the mathematical corpus.

An issue is that the EuDML cooperation and interoperability model *can* scale, but *not that much*. There are around the world a number of institutions that act as local DMLs and would be happy to join a global DML project in the lines of EuDML. However, the well-managed long-standing and scientifically reliable digital libraries that we expect to partner with are not many, and do not host more than a third of the whole mathematical corpus. This is already a lot, and it would certainly be very useful to reach this number, but it is still far from the grand vision of an exhaustive library, or even of a one-stop shop to the established treasures of the mathematical literature.

From the experience of the contacts we had during this project, we can tell that the math publishing landscape is quite diverse, with quite diverging stakes, and it seems difficult to accommodate with all views and policies. Especially, there is a strong attraction of concentration in the digital realm, which opposes to the creation of an alternative archive for universal open access. Moreover, we have to convince publishers with very different profiles that cooperation is fruitful to them (although the same kind of cooperation certainly doesn't mean the same to very different kinds of publishers).

This said, it would probably be possible to go from 30% to 60% by allowing much more volatile content to be found at places such as open archives, personal home pages, community repositories. Methods and technology to index and keep track of that kind of Web content (and hopefully, to keep track of its scientific validation status as well) are very different from what was experimented in this project. It would rather rely on Web crawling and could be tuned by publishing simple standards to push keywords or hints that the posted content is part of this loose version of the DML. This could in fact be tested as a separate endeavour, and if it succeeds, could become one DML content provider. Just for the case of France I tried to list the various sources of DML-related material in [3], which is a long list, although the material that is likely to end up in a system like EuDML is limited to that from NUMDAM [21], Gallica-Math [18] (which is already there), TEL [27], and Gallica books [17].

Another mean to enlarge DML content is digitisation. Public domain (or decades old, out of print) books have probably the biggest potential as a still relevant huge mathematical knowledge reservoir.

4.2. **Technical challenges.** We still lack a full digitisation work-flow starting from paper (or some flavour of low-grade digital file) and outputting usable structured metadata and full-text enabling mathematical knowledge mining, interlinking, etc. Even routine operations such as starting from a scanned journal volume

and ending up with a database of articles with reasonably accurate metadata and extracted full-text are not available right away without a lot of software development or human input.
We still face a fine-grain metadata shortage that won't be overcome through manual work because of the volume to handle. We should engage in all strategies to incrementally enhance metadata or metadata-generating technology. Of course a lot of this is not math-specific but we have to be able to capture and store non-textual metadata everywhere.

Conclusion

The EuDML project has assembled a corpus of 225,000 mathematical documents, which are now much easier to find and navigate, and much better integrated within the Web. As their metadata were converted and stored in a homogeneous format, it will be also much easier to deal with that content in the future, and link it with existing or future infrastructures relevant for the retrieval of scientific reference literature.
A number of tools were developed, deployed, and partially evaluated (from basic aggregation to accessible math through math-mining and deep interlinking).
Finally, the experience of the first cross-repository, trans-national DML effort shows that it can be done, and it can go on with new partners and the support of the scientific community.

Nevertheless, this is only a promising début A number of threats counter the objective of archiving the whole mathematical corpus in a modernized free public digital library. Going further will need a lot of energy and resources from all over the world. The question whether the needed resources to achieve this goal will be allocated is open.

References

[1] Joseph Baker et al., *Maxtract, a tool for converting PDF into formats such as LaTeX, MathML and text*, `http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/maxtract.php`.

[2] Thierry Bouche, *Some thoughts on the near-future digital mathematics library*, Towards Digital Mathematics Library. DML 2008 workshop, Birmingham, UK, July 27th 2008 (Brno) (Petr Sojka, ed.), Masaryk University, 2008, `http://eudml.org/doc/221606`, pp. 3–15.

[3] Thierry Bouche, *Report on the current state of the french dmls*, Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009, Masaryk University Press, 2009, http://eudml.org/doc/220779, pp. 61–70.
[4] Thierry Bouche, *Digital Mathematics Libraries: The Good, the Bad, the Ugly*, Mathematics in Computer Science **3** (2010), no. 3, 227–241, Special issue on Authoring, Digitalization and Management of Mathematical Knowledge (Serge Autexier, Petr Sojka, and Masakazu Suzuki eds.).
[5] Davide Cervone et al., *MathJax, Beautiful math in all browsers*, http://www.mathjax.org/.
[6] Committee on Electronic Information Communication of the International Mathematical Union, *Best Current Practices: Recommendations on Electronic Information Communication*, Notices of the AMS **49** (2002), no. 8, 922–925, http://www.ams.org/notices/200208/comm-practices.pdf.
[7] ———, *Digital Mathematics Library: A Vision for the Future*, http://www.mathunion.org/ceic/recommendations-publications/, 08 2006.
[8] *Free access to archived articles of primary mathematics journals*, http://www.elsevier.com/physical-sciences/mathematics/archived-articles.
[9] *EuDML: The European Digital Mathematics Library*, http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503.
[10] *EuDML oai-pmh server*, 2013, https://eudml.org/oai/OAIHandler?verb=Identify.
[11] *EuDML Opensearch*, 2013, https://project.eudml.org/opensearch.
[12] *EuDML REST Services*, https://project.eudml.org/rest-services.
[13] *EuDML metadata schema specification (v2.0 – final)*, 2013, https://project.eudml.org/eudml-metadata-schema-specification-v20-final.
[14] *EuDML metadata schema validation tool*, 2013, http://eudml.mathdoc.fr/eudml-validation-demo/.
[15] *EuDML tools & demonstrations*, 2013, https://project.eudml.org/tools-technical-specifications.
[16] *Europeana Technical Requirements: Europeana Semantic Elements specifications*, 2012, http://pro.europeana.eu/technical-requirements.
[17] *Gallica, the digital library of the Bibliothèque nationale de France*, http://gallica.bnf.fr/.
[18] *Gallica-Math, a front-end to some mathematical content from Gallica*, http://portail.mathdoc.fr/GALLICA/.
[19] José Grimm, *Tralics: a LaTeX to XML translator*, http://www-sop.inria.fr/marelle/tralics/.
[20] U.S. National Library of Medicine National Center for Biotechnology Information, *Journal archiving and interchange tag library, niso jats version 1.0*, August 2012, Full online documentation at http://jats.nlm.nih.gov/1.0/.
[21] *NUMDAM, Numérisation de documents anciens mathématiques*, http://www.numdam.org/?lang=en.
[22] Andrew Odlyzko, *Open access, library and publisher competition, and the evolution of the general commerce*, http://arxiv.org/abs/1302.1105.
[23] Radim Řehůřek and Petr Sojka, *Gensim, Semantic Similarity of Text Documents*, https://mir.fi.muni.cz/gensim/index.html.

[24] David Ruddy, *The evolving digital mathematics network*, Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009, Masaryk University Press, 2009, http://eudml.org/doc/220643, pp. 3–16 (eng).
[25] Petr Sojka and Martin Líška, *WebMIaS, Math Indexer and Searcher*, https://mir.fi.muni.cz/mias/.
[26] Masakazu Suzuki et al., *InftyReader, a software to recognize scientific documents including mathematical expressions*, http://www.inftyproject.org/en/index.html.
[27] *TEL, Thèses en ligne*, http://tel.archives-ouvertes.fr/.

Received February 23, 2013

Cellule Mathdoc (UMS 5638), Université Joseph-Fourier (Grenoble 1), B.P. 74, 38402 Saint-Martin d'Hères
E-mail address: thierry.bouche@ujf-grenoble.fr